Toward an Interagency Language Roundtable Based Assessment of Speech-to-Speech Translation Capabilities

Douglas Jones, Wade Shen, Brian Delaney, Martha Herzog,
MIT Lincoln Laboratory
Lexington, MA 02420, USA
{DAJ,BDelaney,SWade}LL.MIT.edu
MHerzog2005@comcast.net*

Michael Emonts, Sabine Atwell , James Dirgin, Neil Granoien, Sargon Jabri, Jurgen Sottung,
DLI Foreign Language Center
Monterey, CA 93944, USA
{Michael.Emonts,Sabine.Atwell,James.Dirgin,
Sargon.Jabri,Jurgen.Sottung}
@monterey.army.mil; granoien@sbcglobal.net*

Timothy Anderson,
Air Force Research Laboratory
Wright Patterson AFB, OH 45433, USA
Timothy.Anderson@wpafb.af.mil

Timothy Hunter
United States Army Intelligence Center
Ft. Huachuca, AZ 85613, USA
Timmie.Hunter@us.army.mil

Abstract

We present observations from three exercises designed to map the effective listening and speaking skills of an operator of a speech-to-speech translation system (S2S) to the Interagency Language Roundtable (ILR) scale. Such a mapping is non-trivial, but will be useful for government and military decision makers in managing expectations of S2S technology. We observed domain-dependent S2S capabilities in the ILR range of Level 0+ to Level 1, and interactive text-based machine translation in the Level 3 range.

1 Introduction

We present observations of three exercises that we conducted at a three-day workshop in September 2005 at the Defense Language Institute Foreign Language Center in Monterey, California. The purpose of this workshop was to identify which ILR based testing methods might be suitable for adaptation to S2S evaluation. The U.S. Department of Defense uses ILR-based testing for high-stakes tests that assess foreign language skills on the part of human language learners. Being able to characterize S2S technology in terms of the ILR will be useful for government and military decision makers in relating current S2S capabilities and limitations to a known measure. As expected, we learned that the current methods of testing foreign language learners need to modified, in some cases quite substantially. The primary difference for S2S evaluation is that domain dependence required us to make adjustments. Translation errors were also a factor.

Our starting point for measuring S2S capabilities was the Oral Proficiency Interview (OPI), a standardized test of speaking skills that is also able to test listening skills. However, the OPI is designed as a general language proficiency test, not a domain-specific or job-related performance test. Our focus here is on a job-related test, since S2S systems are currently domain-dependent. Thus we needed to create a more specialized test, which we are calling the Job-related Speaking Performance Test (JSPT) that can also be used to test listening skills.

• This work is sponsored by the Defense Language Institute under Air Force contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.
* Dr. Herzog is working under Research Subcontract with MIT Lincoln Laboratory.
* Dr. Granoien retired from DLI in April 2006 and continues to be active in language learning research.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
AUG 2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Toward an Interagency Language Roundtable Based Assessment of Speech-to-Speech Translation Capabilities	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,LIncoln Laboratory,244 Wood Street,Lexington,MA,02420**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
AMTA 2006, Boston, MA, August 8, 2006. U.S. Government or Federal Rights License

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

We first had to address several challenges for constructing a realistic JSPT dialog vignette using an S2S system, in this case, a medical interview conducted in English and in Mandarin Chinese. We discuss details below, but the conclusion of the first exercise is that we were not able to observe speaking and listening skills that exceeded Level 0+ and generally felt that the participants were unable to communicate effectively.

In the second exercise, also using the medical JSPT format, we were able to observe speaking skills that may reach Level 1 and listening skills that may exceed 0+. However, since the speaking and listening skills are assisted by machine, and are not tightly coupled, as they are in a language learner's mind, it is no longer obvious how to combine these numbers in a way that predicts task performance. This point is subtle but important: when language learners speak an utterance, we assume that they can understand that same utterance if someone else says it. However, this assumption is no longer valid for a person using an S2S system. Consequently, we expect to need to report an entire array of numbers, some based on ILR skills, and others based on system-internal measures of performance, such as word-error rates and translation error rates according to various commonly accepted scales, such as BLEU (Papineni, et al. 2001). Moreover, we will not be able to provide a strong relationship between this array of numbers and ILR skills until we have been able to validate predictions of the ability of the user to accomplish various tasks, which have themselves been rated according to ILR difficulty.

In the third exercise, we were able to conduct one OPI using a text-to-text (T2T) system that followed the standard OPI format up to Level 3. By removing the errorful speech recognition component, and by allowing the interviewer and examinee to communicate interactively in their native languages, one interview was conducted following the general OPI guidelines. The questions and answers were typed into the text-based MT system for translation, so we note that the interview was no longer specifically oral. We will discuss the details below.

ILR Skill Level	Interactive Comprehension	Global Tasks and Functions
Memorized Proficiency (0+)	The individual understands a number of short, memorized utterances in areas of immediate needs; frequent, long pauses and repeated requests for repetition.	Can make statements and ask questions using memorized material.
Elementary Proficiency (1)	A native speaker must often use slowed speech, repetition, paraphrase, or any of these to be understood by this individual. Misunderstandings are frequent, but the individual is able to ask for help and to verify comprehension of native speech in face-to-face interaction.	Can create sentences; begin, maintain, and close short conversations by asking and answering simple questions; satisfy simple daily needs.
Limited Working Proficiency (2)	The individual can get the gist of most everyday conversations, but has some difficulty understanding native speakers in situations that require a specialized or sophisticated knowledge. (May require a native speaker to adjust to his/her limitations in some way).	Can describe people, places, and things; narrate current, past, and future activities in full paragraphs; state facts; give instructions or directions; ask and answer questions in the work place; deal with non-routine daily situations.
General Professional Proficiency (3)	In face-to-face conversation with natives speaking the standard dialect at a normal rate of speech, comprehension is quite complete. Although cultural references, proverbs, and the implications of nuances and idiom may not be fully understood, the individual can easily make repairs.	Can converse extensively in formal and informal situations; discuss abstract topics; support opinions; hypothesize; deal with unfamiliar topics and situations; clarify points.
Advanced Professional Proficiency (4)	Can understand native speakers of the standard and other major dialects in essentially any face-to-face interaction. Can understand the details and ramifications of concepts that are culturally or conceptually different from his/her own. Understands shifts of both subject matter and tone.	Can tailor language to fit the audience; counsel; persuade; represent an official point of view; negotiate; advocate a position at length; interpret informally.
Functionally Native Proficiency (5)	(No gaps in comprehension, including all details and nuances.)	Functionally equivalent to a highly articulate, well-educated native speaker.

Figure 1: Excerpt from OPI Rating Factor Grid

2 ILR Measurements for Speaking and Listening Skills

The Oral Proficiency Interview (OPI) is a standardized, high-stakes test that is administered to U.S. Government personnel to measure their speaking skills and that can also be used to measure their listening skills. The OPI is designed to provide a reliable answer to the simple question of how well a person speaks a foreign language. It is a carefully structured interview that has been validated by decades of use in the U.S. Defense Department, the Foreign Service Institute in the State Department, and other organizations.

The OPI is designed to assess general foreign language skills and is not tailored to measure specialized domains. Figure 1 on the previous page shows a brief description of the ILR skills assessed in the OPI (DLI English Language Center, 2006). More complete descriptions are available (ILR web site, 2006).

An ordinary OPI testing up to Level 3 should take no more than half an hour, beginning with a warm-up phase, checks for speaking levels, probes for ability to sustain increased difficulty levels, and a short wind-down phase; see the OPI manual for details (DLI Foreign Language Center, 1999). We expect interviews applied to S2S technology to take somewhat more time, allowing for system latency, etc.

During our workshop, we had access to two Mandarin Chinese / English S2S systems developed for the medical domain. Because of the domain limitation, we were not able to conduct a generic OPI, as should be obvious in examining the skill descriptions in Figure 1. Current S2S systems are typically designed to be domain-dependent in order to reduce speech recognition and translation errors.

V. PHRASES FOR THE DOCTOR
A. Case History

180. I am ……, and I will look after you.
181. When were you taken ill?
182. When were you wounded?
183. When did you have the accident?
184. Please show me on the watch/calendar, the time/date.
185. Have you been sick/wounded previously?
186. What kind of disease(s)/wound(s) (470-485, 500-571)?
187. Show me where.

Figure 2: Sample Phrases from Army Pamphlet 40-3

Rather than attempting to administer an OPI using the S2S devices, we constructed a plausible medical scenario with portions of a dialog occurring at increasing levels of difficulty as measured by the ILR speaking and listening skill levels.

2.1 A Level 1 Medical Job-Related Speaking Performance Test

The dialog for the medical interview was constructed using specific phrases drawn from the Medical Service Multilingual Phrase Book, Dept. of the Army Pamphlet 40-3, an official military medical interview manual (Department of the Army, 1971) which lists Medical Equivalent Words and Phrases. Some sample phrases from DA Pam 40-3 are shown in Figure 2. Twenty of the thirty dialog turns we outlined were based on phrases from DA Pam 40-3. The reason for using an official manual is to illustrate a conceptual principle, namely, the idea of working from official training and doctrine in constructing dialogs for other domains of interest. By creating a tight linkage between personnel training and dialog evaluation, we hope to create both scientific and programmatic links between technology development and requirements for defense department needs.

In order to provide a rationale for the general linguistic features we needed to test in our dialog, we specified the context for our simulated interview as follows: It takes place in a casualty support hospital in a war zone. The patient is indigenous personnel. She was walking with her husband and her teenage daughter, and she fainted on the street. She has abdominal pain. She has a ruptured appendix, but all she knows is that she has pain in her lower right quadrant. She is separated from her family and wants to know where her husband and child are. She is a trained nurse, therefore an informed patient. The reason for creating an informed patient role was so that it would be natural for the patient to produce more complex text in the dialog, especially at Level 2. The reason for assigning the gender roles as we did was to create a culturally plausible reason why the patient would not want to be examined by a strange doctor without her spouse present. The patient's reserva-

tions stressed the interaction in an unexpected way, as we discuss below.

Figure 3 shows a synopsis of the dialog we constructed. The person who played the doctor followed the script fairly closely, in part to facilitate easier coding of the experimental results. The patient was provided the short list of basic facts but did not have a precise script. The dialog increases in difficulty as it progresses, as is clear from the synopsis.

The bulk of the 30-turn dialog occurs at Level 1. Overall, the doctor uses slightly more complex language, whereas the patient's language is roughly split between Levels 0+ and 1. Both doctor and patient have a few turns of Level 2 dialog. The specific breakdown of the doctor's turns: no turns at Level 0+; 28 turns at Level 1; 2 turns at Level 2. For the patient: 12 at Level 0+; 13 at Level 1; 5 at Level 2. The primary purpose of the Level 2 material is to test whether the system tops out at Level 1. In the future, when S2S systems enable easy negotiation of this dialog, we would extend the difficulty to probe higher levels.

The overall performance on the first medical interview is shown in Figure 5. Of the 30 utterances spoken by the doctor, there were 20 failures, shown as 67% Utterance Failure Rate, using the negative direction for contrast with the success rate. For the patient, there were 46 failed utterances, more than the total number of 30 turns because of multiple failures. Any given turn was judged successful if it was completed with two or fewer failures for each side of the conversation. 6 of the 30 turns were successful, shown as 20% Turn Success Rate. The overall observation is clear: The conversational participants are failing to communicate with the S2S device at Level 1, since only one turn in five is successful. As we mentioned earlier, this system was an early research prototype. We repeated the exercise with a more recent system, with markedly better results.

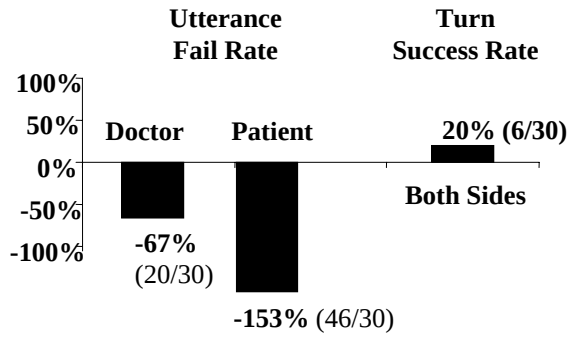

Figure 5: Medical JSPT Results for First Interview

Part I. Registration, performed by orderly.
[1] What is your name? [2] What is your address? [3] What is your phone number? [4] What is your date of birth? [5] Are you married? [6] What is your husband's name? (next of kin). [7] Do you have children? [8] What is your religion?

Part II. Interview, performed by doctor:
[9] Good morning, I am Dr. ….. I work in this hospital, and will take care of you. [10] Are you ill? *(yes).* [11] How long have you felt ill? *(2 days).* [12] Do you feel pain? *(yes)* [13] Tell me where it hurts *(has a pain in abdomen)* [15] Have you taken any medicine? *(no)* [16] Are you allergic to any medication? *(no)* [17] What did you eat for the last two days? *(says she hasn't eaten.)* [18] What other symptoms do you have? [19] Can you describe exactly what happened to you? *(tells story)* [20] Has anyone else in your family been affected? [21] How did you come to this hospital? *(says she fainted)* [22] What do you remember? *(doesn't remember because she fainted)* [24] Do you mind if I examine you? *(yes, the patient minds)* [25] I believe you may have appendicitis [26] We need to do some tests. [27] Doctor gives detailed explanation of tests. [28] Patient expresses concerns. [29] Doctor gives detailed explanation of need to operate [30] Closing remarks by both doctor and patient.

Figure 3: Synopsis of Medical JSPT

We repeated the exercise with the same dialog and the same native speakers of English and Mandarin Chinese. The overall performance on the second medical interview is shown in Figure 4 and is around three times better than before. Of the 30 utterances spoken by the doctor, only 7 failed, shown as 23% utterance failure rate. For the patient, 19 of 30 failed, i.e., 63%. As before, each turn was judged successful if it was completed with two or fewer failures for each side of the conversation. 22 of the 30 turns were successful, shown as 73% below.

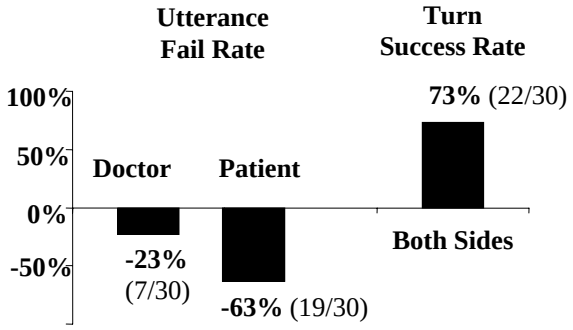

Figure 4: Medical JSPT Results for Second Interview

Caveats and Observations for Medical JSPT

Despite the marked increase in performance, some caveats are in order. In addition to noting whether the information in the turns was communicated successfully, we made an informal notation as to whether the turn was translated in an awkward fashion. In some cases, there were odd word choices or minor defects in the translation, but these were in general the type of error that may be overcome by a person who is accustomed to interacting with foreign speakers. For example, in turn 4 (Figure 3: Synopsis of Medical JSPT), the doctor asks "What is your date of birth?", to which the patient replies "六六年 五月 十二号, recognized as "6六年 五月 十二号" and translated as *"6 six years may twelve."* The doctor said he understood this to mean May 12, 1966, which was in fact the correct answer.

We do not necessarily expect that people who are unaccustomed to communicating with foreigners would be able to negotiate the awkward translations so successfully, and we also do not expect that every educated guess at the intent of the speaker would be correct. For the doctor there were 5 awkward translations and for the patient 13. In other words, 17% of the doctor's translations were awkward, as was 43% of the patient's. The patient also had one misleading translation, one partial translation, and one excessively verbose turn (i.e., 3% for each of those types). We may well expect that with less well-trained operators, the awkward translations would drastically degrade their ability to communicate.

To give a sense of the misleading, partial and verbose translations, we turn to the following examples, which refer to the dialog turns shown in Figure 3: Synopsis of Medical JSPT.

The misleading translation occurred in turn 24 when the doctor asked the patient if it was OK to examine her, and the MT reply was "i think i gentleman". The doctor assumed that the patient said he was "being a gentleman" by asking very carefully about the exam. The patient was actually asking for the doctor to wait for the patient's husband, which was translated as 'gentleman' (先生). In an actual examination, the confusion would have soon become apparent and embarrassing.

The partial translation was the correct recognition of six of seven digits in the phone number for turn 3.

The verbose translation was an answer to the doctor's question in turn 21: "How did you come to this hospital?", to which the patient replied that she did not know, which was correctly translated, but continued with additional remarks that were mistranslated. The doctor was confused by the largely garbled response and needed to re-state his question. That turn was not successful.

The overall trend is clear: the dialog is marginally successfully, albeit awkward, and occurs primarily at Level 1. Turn-by-turn coding for the second interview is shown in the Appendix.

Extending the Level 2 part of the dialog to have more turns would increase our confidence that the trend is robust, as would repeating the dialog using more test subjects as role players.

2.2 Simulated German Oral Proficiency Interview

In the third exercise, we conducted an OPI using a text-to-text (T2T) system and followed the standard OPI format up to Level 3. The interview took about two hours. We felt that we could have continued and probed higher levels but did not because of time considerations. We were able to successfully conduct the interview by removing the errorful speech recognition component. The interviewer and interviewee communicated verbally in their native languages while others typed their input into the T2T system and read the MT output. This interview was conducted following the general OPI guidelines, although we note that the interview was no longer purely oral due to the way we used the technology.

We also point out that it was conducted between parties who are accustomed to garbled communication, either on the part of human language learners or on the part of errors in machine translation technology. Though informal, we took care to maintain the integrity of the experiment. For example, we did not allow the participants to try to read each other's native input to the MT system during the exercise. They could, however, hear each other, just as in a real OPI.

One overall remark is that the participants in the interview were able to compensate for mistakes in the MT system by re-phrasing questions and an-

swers, and by being forgiving of errors in translation where they did not impede the flow of information.

Selections from the beginning of the interview are shown in Figure 6. Each turn is numbered as it occurred in the actual interview. Bold turns indicate questions posed by the interview. Responses by the examinee are shown in italics. Notice that first turn is error-free: Mrs. Schmidt introduces herself and says "Good Day". John gives his name and says "Hello". In the second turn, the examinee compensates for errors and makes the reasonable and correct educated guess that the question about surnames is a prompt for his own surname, which he gives as "Cooper".

1. **Good day. I am Mrs. Schmidt.**
   Hello. My name is John.
2. **And like hei ss EN it with surnames**
   My last name is Cooper.
3. **Mr. Cooper, how are you today?**
   I am very well today. How are you?
4. **Mr. Cooper, are you already for a long time in California**
   Um. I have been here since Monday.
5. **Where are you ago?**
   I came here from Chicago.
6. **Did you grow up in Chicago?**
   No, I am from New Jersey.
7. **Then you went also into New Jersey to the school**
   I went to high school in New Jersey, but I was in Pasadena for college.
8. **And do you mean the university with "university "?**
   I don't understand.
9. **Did you go into Pasadena to the university?**
   I went to school at CalTech. I received both graduate and undergraduate degrees there

….

Figure 6: Beginning of OPI (Examinee Perspective)

Turn 3 proceeds without error. Turn 4, which we are viewing from the English examinee's point of view, makes it sound like the interviewer is the one with a foreign accent: the German sentence pattern is conveyed with English words. The difficulty in turn 7 reflects a cultural difference as much as a linguistic difficulty. The interviewer asks whether the examinee went to 'school' in New Jersey, and he replies that he went to high school in New Jersey, but went to college in Pasadena. His translated reply is confusing because it contains both the English term "High School" and the German 'Hochschule'. See the German interviewer's point of view in Figure 7.

7. **Dann sind Sie auch in New Jersey zur Schule gegangen**
   Ich ging zur High School in New Jersey, aber ich war in Pasadena für Hochschule.
8. **Und meinen Sie mit "hochschule" die Universitaet?**
   Ich verstehe nicht
9. **Sind Sie in Pasadena zur Universitaet gegangen?**
   Ich ging zur Schule am CalTech. Ich empfing beide graduierten und Nichtgraduiertgrad dort.

Figure 7: Begin German OPI (Interviewer Perspective)

What is of interest is that the participants are able to overcome the error in translation: the interviewer asks the question in a different way in turn 9, and the examinee gives an appropriate response.

After the warm-up and Level 1 phase of the interview, the interviewer directs the discussion to more complicated material. In Figure 8, she asks the examinee to describe and give directions to a favorite restaurant that he likes to visit with his girlfriend. As is evident in turns 30-33, the examinee was able to give directions, which is considered a Level 2 task.

30. **Sometimes Mr. Cooper, you go out with your friend to the meal?**
    Yes, we both like Chinese food very much. In fact, she is one quarter Chinese.
31. **Is the Chinese restaurant far away from your house or your dwelling?**
    The restaurant is only a mile away from my house, so we can go whenever we like.
32. **Mr. Cooper, you can say to me times exactly, how one comes from your house to this restaurant?**
    First you take a right on to William street and then go straight until Washington street. When you get to Washington street make a right. Then go past the Staples store and make a right into the parking lot. The restaurant is in the back. It is called China Café.
33. **Mr. Cooper, Sie said, you often go into this restaurant. Can you me times the restaurant from the inside describe. How does it look?**
    It is nothing fancy. There is a buffet on the right, but I usually order from the menu. There are booths along the walls and tables in the middle. The walls are red and there are Chinese decorations throughout.

Figure 8: Level 2 Section of German OPI

The interviewer kept increasing the difficulty of the interview, moving on to questions about the examinee's college experiences: what kind of university did he attend, what were the admission criteria, and questions of this nature.

Once it became clear that they could communicate at the factual level, at times making repairs to the conversation by adapting to the idiosyncrasies of the translation system, the interviewer moved on to more difficult questions that involve higher level reasoning and more nuanced implications. She moved on from factual details about the examinee's personal college experiences to more general questions about the role of government in establishing educational policy.

40. Mr. Cooper. I have now times another question. How do you explain yourselves it that so many young Americans do not decide for a scientific career, and that this fact has serious consequences for the American economy?
I do not understand completely myself. I have always enjoyed math and science, but I think Americans in general do not value these subjects. It seems that other countries do a much better job at recruiting science students. For example, the majority of engineering students in American universities are not from the US.

41. Yes, I already understand. But on which educate-political reasons you lead back that?
I think it has to do with the funding of science programs by American politicians. The space program in the early 1970s really helped encourage many students to study math and science. Today, there is very little exposure to these subjects in the media.

42. What for measures would have to be met on the part of the government, in order to change and improve this condition?
What we really need is more funding at all levels of education as well as federal funding to provide challenging jobs for graduating students. If a student does not believe that a career path as available after studying science, why would that student go and get an advanced degree in the sciences.

Figure 9: Level 3 Section of German OPI

As is evident in Figure 9, the examinee is able to argue for additional funding for education and for government advocacy for career paths in science. It is interesting to see that when the examinee says "I do not understand completely" in turn 40, he is referring to the complex political situation, not the MT output. The participants in the interview were felt to be communicating reasonably well at Level 3.

3 Conclusions

In the first exercise, using an early S2S research prototype for the medical interview, we observed that the conversational participants essentially failed to communicate. In that primarily Level 1 dialog, only one turn in five was successful. When we repeated the exercise with a more recent system, we observed markedly better results: the dialog is marginally successful. We do not necessarily expect that people unaccustomed to garbling of their native language would do so well. A remarkable outcome of the German test is that, despite the necessity to type in the language and the long pauses that followed, it worked very well as a test of general language. We may have expected a domain-specific test to work this well but not a test following most of the conventions of a general proficiency test. Of course, German and English are closely-related languages. In future work we will test other languages as well.

4 Future Work

As a caveat, we can convey our current levels of understanding with the terms "Validated" and "Pre-validated" ILR-based experiments. A Pre-validated ILR-based experiment for S2S, such as what we present here, is designed to measure ILR speaking and listening skills, and follows conventional ILR testing methodology as closely as possible. However, a validated ILR-based experiment will have been validated in experiments demonstrating statistical reliability with human subjects accomplishing a variety of tasks that have been previously rated at traditional ILR levels. We recognize that substantial work needs to be done to define pre-validated requirements well enough to enable the move to the validated results we desire.

We also wish to relate our emerging ILR-based evaluation measures to other types of work, such as the research involving model patient dialogs (Belvin et al. 2004), and the DARPA TransTAC evaluation methods (Hsiao et al. 2006).

Our next steps are to design a suite of tests around specific tasks that language learners at lower levels are able to accomplish. Since each S2S system is constructed for a specific domain, each domain needs to have its own test of tests for tasks that can be accomplished in that domain. Moreover, we need to design a common framework as a bridge to conventional ILR testing, in order to measure the abilities of people without foreign language skills working with the assistance of S2S devices. We would consider the tests to be validated when they have been run with enough subjects to reliably predict task performance. Eventually we would expect to construct a mosaic of validated experiments on a domain-by-domain basis that converges to correspond with general foreign language capabilities exhibited by people.

References

Robert Belvin, et al. Creation of a doctor-patient dialogue corpus using standardized patients. In Proc. LREC, Lisbon, Portugal, 2004.

Ray Clifford, et al. 2004. "The Effect of Text Difficulty on Machine Translation Performance -- A Pilot Study with ILR-Rated texts in Spanish, Farsi, Arabic, Russian and Korean". *Proceedings of Language Resources and Evaluation Conference*, Lisbon.

Defense Language Institute Foreign Language Center. 1999. *OPI 2000 Tester Certification Workshop Training Manual for Tester Certification*.

DLI English Language Center. 2006. *OPI Rating Factor Grid*, from English Language Training Support for Security Assistance Offices; Lackland AFB, TX.

Hsiao, Roger, et al. 2006. "Optimizing Components for Handheld Two-way Speech Translation for an English-Iraqi Arabic System". ICSLP 2006.

Interagency Language Roundtable Website. 2006. ILR Language Skill Level Descriptions. http://www.govtilr.org

James R. Child, Ray T. Clifford and Pardee Lowe, Jr. 1993. "Proficiency and Performance in Language Testing". *Applied Language Learning*, Vol. 4.

Department of the Army. 1971. DA Pamphlet 40-3: *Medical Service Multilingual Phrase Book*, NAVMED P-5104/APF 160-28, dated 31 May 1971.

Douglas Jones, et al. 2005. Measuring Human Readability of Machine Generated Text: Three Case Studies in Speech Recognition and Machine Translation". *HLT Special Session*, ICASSP 2005.

Kishore Papineni, et al. 2001. "Bleu: a Method for Automatic Evaluation of Machine Translation" *IBM Computer Science Research Report RC22176* http://domino.watson.ibm.com/library.

Appendix: Medical Job-Related Speaking Performance Test (JSPT) Score Sheet for Second Interview

Turn	DA Pam 40-3	ILR Level		N Fails		Pass	Awk
		Dr	Pt	Dr	Pt		
1. NAME	51	1	0+	0	2	0	1
2. ADDRESS	64	1	0+	0	2	0	1
3. PHONE		1	0+	0	2	0	1
4. DOB	64	1	0+	0	1	1	1
5. MARITAL		1	0+	0	0	1	0
6. KIN		1	0+	0	1	1	1
7. CHILDREN		1	0+	0	0	1	0
8. RELIGION	55	1	0+	1	0	1	1
9. GREETING	6; 180	1	0+	0	0	1	0
10. ILL	10	1	0+	1	0	1	1
11. DURATION	25	1	0+	1	0	1	1
12. PAIN	80	1	0+	0	0	1	1
13. LOCUS	187	1	1	0	0	1	1
14. WATER		1	1	0	0	1	1
15. MEDICINE	212	1	1	0	0	1	1
16. ALLERGIES	213	1	1	0	1	1	1
17. FOOD	202; 17	1	1	1	0	1	1
18. SYMPTOMS		1	1	1	2	0	1
19. EVENT	208	1	2	0	0	1	1
20. FAMILY		1	1	1	0	1	1
21. ARRIVAL		1	1	0	2	0	1
22. REMEMBER	208	1	2	0	0	1	0
23. STORY	250	1	2	0	2	0	1
24. EXAMINE	250	1	2	0	2	0	1
25. APPENDIX	290; 507	1	1	0	0	1	0
26. TESTS	269	1	1	0	0	1	0
27. EXPLAIN	290; 507; 269; 293; 298	2	1	0	0	1	0
28. CONCERNS		1	1	1	0	1	1
29. OPERATE	293	1	1	0	2	0	1
30. CLOSING		2	2	0	0	1	1